KISTI의 생명정보 인프라 구축

손 현 석 (KISTI 바이오인포매틱스 센터장)

생명정보 인프라 구축을 통해 바이오인포매틱스 핵심기술을 축적함으로써 생명정보 전반의 기반기술을 구축할 것이다.
유수 생명정보 데이터베이스에 대한 향상된 검색기능을 제공하고 새로운 개념의 통합검색 관리시스템을 구축함으로써 관련 연구자들의 연구능력을 고취함과 동시에 독창적인 검색시스템으로 보다 풍부한 연구기회를 부여하여 국내 생명정보 연구에 기여할 것이다.

생명정보 인프라 구축은 생명정보 관련 연구개발의 기반이 되는 대용량의 정보를 효율적으로 저장 관리할 수 있는 데이터베이스 구축, 서열비교, 분자 비교 가시화, 검색 처리 및 분석 시스템을 구축 등을 포함하며 그 중요성이 점차 커지고 있다.

21세기에 들어서면서 기하급수적으로 증가하고 있는 유전정보와 생물학적 데이터 등 생명정보 연구의 필요성이 급증하고 있으나, 대부분의 생명공학 관련 산·학·연 연구자들은 연구 분야에 적합한 IT 기술을 적용할 수 있는 적절한 방법을 보유하고 있지 못할 뿐 아니라, 급속도로 생산되고 있는 생명정보들을 수집·분석·관리할 수 있는 기반환경이 갖추어지지 못한 것이 현실이다. 따라서, 생명공학 관련 연구자들의 요구에 부응하는 정보를 효율적으로 제공하기 위해 IT를 기반으로 한 정보수집과 도구 및 요소기술 개발 등을 통한 정보공유 및 공동연구의 생명정보 인프라 구축이 절실히 요구된다.

최근 가장 각광받는 분야인 바이오인포매틱스는 의약, 화학, 농업, 환경에 이르는 다양한 분야에 영향을 미치고 있고, 향후 관련기술의 발전과 더불어 여러 성과들이 가시화될 전망이며 연평균 50%대의 고성장으로 거대시장을 형성할 전망이다. 세계 바이오시장의 경우 2000년 540억 달러에서 2013년 2,100억 달러로 연평균 11%의 높은 성장이 이루어질 것으로 예상되고 있다. 따라서, BT 외에도 생명정보 인프라 구축을 위한 기반이 되는 IT 기술의 개발 및 이를 바탕으로 한 생명정보 검색 분석 시스템 개발 및 지원·활용체제 구축이 필요하다.

본 연구사업을 통해 생명정보 관련 연구자들에게 필요한 기능들을 제공하기 위해 분산되어 있는 생명정보 데이터를 통합하고 이를 효과적으로 연구에 활용할 수 있도록 생명정보 데이터베이스의 안정적인 최신 정보 유지와 다양한 유전체 사업을 통해 얻은 대량의 생명정보에 대한 검색 시스템 개발 및 생명정보의 효율적인 관리와 서비스를 위한 생명정보 기반기술을 확립해 나갈 것이다.

Ⅰ. 연구의 필요성

1. 경제적 중요성

바이오인포매틱스 시장은 향후 거대시장으로 발전될 것이고 세계적으로 계속 확대되고 있는 성장 과도기에 있다. 생물학적 도구로서의 바이오인포매틱스는 생물학 뿐 아니라 의약, 화학, 환경, 농업 등 다양한 분야에서 경제적 효과를 창출해 나가고 있다.

많은 부분 생명정보 연구에 필요한 도구 및 시스템들이 이미 선진국을 중심으로 개발 사용되고 있으며, 기능적인 면에서 지속적인 업그레이드가 유지되므로 국내에서 새로운 도구를 만들고 공유하기에는 많이 뒤쳐져있으나 효과적인 생명정보 인프라 구축을 통해 한계를 극복할 필요가 있다. 국내의 경우, 아직은 선진국에 비해 연구가 미비하나 생명정보 데이터베이스와 분석 소프트웨어를 국내적으로 자체 개발한다면 수입대체 효과를 발생시켜 경제적 이익을 낼 수 있다.

또한, 선진국의 생명정보 이용시 고가의 사용료 지불에 따른 경제적 손실에 대비하여 생명정보 응용연구를 통한 고부가가치 정보를 확보해야한다. IT 기반의 생명정보 인프라 구축 및 바이오인포매틱스 응용 연구 등의 다각적인 생물학적 정보를 연계함으로써 선진국 기업의 생명공학 관련 특허 선점이나 국내기업의 해외 시장 진출의 어려움 혹은 생산기술 저위문제 등을 극복할 수 있다.

IT의 활용은 생명정보의 수집 · 저장 · 분석 · 가공 등을 통해 의약, 화학, 환경, 농업 등의 여러분야로의 부가가치 창출이 가능하다. 신약개발의 경우 생명정보 시스템을 활용한다면 시간과 비용의 대폭 절감이 가능하며, 환경이나 화학 및 생물소재분야의 경우 산업적으로 유용한 생물정보 데이터베이스 및 시스템을 구축하여 활용한다면 새로운 공정과 제품 개발이 가능하다. 또한, 농업분야의 경우 식물 관련 게놈 사업이나 작물 데이터베이스를 구축함으로써 농업의 생산성 및 품질 향상을 가능케 할 수 있다.

IT는 향후 높은 성장을 이룰 것으로 보이는 생명정보산업의 핵심 기반기술로 자리 잡을 것으로 예상되므로 국내에서도 이를 위한 역량을 확보해야한다. 이를 위해서는 생명정보 관련 인프라 구축을 위한 노력이 무엇보다 중요하다. 따라서, 국내실정에 맞는 생명정보 시스템 개발이나 소프트웨어 개발 및 독창적인 생명정보 데이터베이스 구축 등을 통해 국가 경쟁력을 확보해 나가야할 것이다.

2. 사회적 중요성

생명정보 산업은 미래 산업으로서 국내외적으로 많은 관심과 인류의 건강증진이나 질병예방에 기여하는 21세기 사회 · 경제적으로 중요한 역할을 할 것으로 기대된다. 생명정보 연구는 그 범위가 넓을 뿐만 아니라 국내 실정은 아직 초기 정착 단계에 있어 산업계에 이바지할 연구나 정보공유가 미흡한 실정이다. 향후 생명정보를 이용한 유전체 해석 연구나 유전자 치료 기술, 뇌과학 · 뇌광학 연구의 진전, 형질전환 등의 분야에서의 선진국의

급속한 기술 발전으로 선진국과의 기술 격차가 커지게 되면 사회 · 경제적으로 많은 손실을 가져올 것이다. 따라서, 이를 막기 위해서는 BT 뿐 아니라 IT 분야에서 핵심 기반 소프트웨어 개발을 유도하고 이의 보급을 통한 기술 집적도 향상 및 고성능 컴퓨터 활용기술 개발로 생명정보의 통합 데이터베이스 구축과 공동 활용 기반 시스템 구축으로 고부가가치형 생명정보 인프라를 구축해야한다. 또한, 국내 고유의 유전자원 관리 및 활용에 관한 시스템 구축을 통해 기술력 축적과 물적 자원 및 고급인력의 해외 유출을 방지하고, 관련 기관들과의 유기적인 협력 체제를 구축하여 범부처적 협력사업으로 중복투자를 피해 부족한 전문 인력과 원천기술개발에 집중되어야 할 역량이 내부경쟁으로 소모되는 것을 예방해야한다.

본 연구사업을 통해 이러한 문제점들을 해결하고 바이오인포매틱스 분야에서 앞선 선진 외국기업이나 연구소에서 유전자 관련 특허의 선취득으로 인한 의약품 개발 및 특허 사용료 지불 등의 막대한 로얄티 지출과 의료계 종속의 위험 등을 극복하여 국민 보건복지향상에도 기여할 것이다.

3. 기술적 중요성

생명정보는 이미 생물학, 의학, 화학 등 여러 분야의 연구에 많은 영향을 끼치고 있고 필수적 연구 수단이 되었다. 세계적으로 생산되고 있는 대량의 유전정보에 대한 서비스 시스템은 정보의 관리에 주력하여 원활한 사용자 서비스에 애로를 겪고 있고, BLAST 등의 서열비교 소프트웨어는 알고리즘의 복잡도가 높아 기하급수적으로 증가하는 유전자 주석 및 서열 검색을 위한 새로운 방법론이 필요하다.

또한, 국외 데이터의 미러사이트 구축 및 가공에 있어 최신 정보 업 데이트의 어려움이나 특화되지 않은 정보 서비스에 따른 국내 연구자의 서비스 이용 기피 성향이 강하므로, 이를 극복하기위해서는 연구에 직접적으로 필요한 가공된 2차 · 3차 고부가 가치정보 제공이 선행되어야한다. 국내에서 생산되는 데이터도 생성에서 서비스까지의 통합관리와 중복적으로 생산되는 데이터의 공유 혹은 이의 공동 활용체계가 미비하므로 생명정보 인프라를 구축함으로써 이를 극복하는 중요한 기반 기술로 자리 잡을 것이다. 더불어, 본 연구기관과 같은 국가 전문정보 유통기관이 주축이 되어 국내외에 분산되어 있는 연구결과를 통합하고 이를 효과적으로 연구에 적용할 수 있어야 하겠다.

본 연구사업을 통해 유전체 사업의 결과물인 생명정보의 효과적인 처리를 위해 대용량의 유전자 주석에 대한 검색 및 서열검색을 위한 새로운 방법론을 개발하여 생명정보 관련 연구자들에게 최신의 정보와 다양한 연구 방법을 제공할 것이다. 이를 위해 여러 생명정보 데이터베이스간의 상호 연계를 통해 이의 활용방안을 제시하여 생명정보 검색 시스템 개발 및 생명정보 유통 시스템을 구축할 것이다.

II. 연구목표

1. 생명정보 시스템 개발

생명정보의 기반기술이 되는 대용량 생명정보 분산검색을 가능케 하는 주석 검색시스템

을 개발하고, 이미 사용중인 생명정보 검색시스템의 검색속도 및 성능을 향상시키며, 생명정보 검색시스템을 응용한 색인시스템 개발과 단백질 서열정보 검색시스템 및 단백질 서열 분류 시스템 개발을 통해 생명정보 시스템 개발을 주도한다.

또한, 신호전달 경로 데이터베이스 구축 및 네비게이션 시스템 개발을 통해 생명정보 기반기술을 확립한다.

2. 생명정보 활용체제 구축

생명정보 활용체제 구축의 일환으로 생명정보 데이터베이스 구축과 유지 보수 및 자동 업 데이트 모듈을 적용하고, 데이터베이스와 인프라 사용자들의 신뢰성과 편의성을 높여주는 다양한 인터페이스를 개발한다. 사용자 편의를 위한 생명정보 웹 서비스 체제를 위해 UDDI 레지스트리 구축과 생명정보 서비스 등록 및 서비스 검색 시스템을 구축하고, 생명정보 분류체계 작성과 통합·검색·분석을 위한 포탈 서비스 시범 사이트를 구축하고자 한다. 또한, 등록되는 서비스와 사용되는 서비스를 분석하고 생명정보 연구에 활용하기 위한 서비스 분석 및 연구동향 파악을 통해 전문화된 서비스를 제공하기 위해 고부가 생명정보 데이터베이스 구축 및 생명정보 웹 서비스 체제를 구축한다.

더불어, 지속적인 국내외 생명정보 데이터베이스 구축 및 국제적으로 제공되는 데이터에 대한 미러사이트 운영을 위해 신속하고 정확한 자동 업 데이트 기능과 기 구축된 미러사이트나 국내 생명정보와 관련된 학회 데이터베이스와의 연계 검색을 통해 생명정보 이용자들에게 보다 편리한 인터페이스를 개발한다.

3. 생명정보 지원체제 구축

KISTI의 고유 기능인 국내 과학기술의 기반 인프라 구축의 일환으로 생명정보 지원체제를 구축한다. 또한 기 구축된 생명정보 데이터베이스 및 유전정보 분석 서비스의 유지보수 및 새로운 해외·국내 생명정보 데이터베이스와 유전정보 분석 서비스도 시행한다. 대량의 생명정보데이터 서비스 체제를 구축하고 초고속 네트워크 구축 및 서비스로 생명정보 시스템이 연구자들에게 보다 안정적이고 효율적인 정보를 제공할 수 있는 기반 기술을 구축한다. 또한, 이와 같이 구축된 생명정보 지원체제 시스템을 국가유전체정보센터 인프라와 유기적으로 결합하여 통합 서비스 시스템을 개발한다.

Ⅲ. 주요내용

1. 생명정보 시스템 개발

1) 생물정보 검색시스템(Bio-KRISTAL) 개발
   가. KRISTAL-2002 정보시스템 기반의 생명정보 전용 검색시스템 개발
   나. 자연어 처리 기반의 새로운 생명정보 색인기법을 개발
   다. 유전자 염기서열, 단백질 아미노산서열 등 생명정보 검색에 적합한 새로운 검색 모델 개발

2) 생명정보검색시스템을 응용한 단백질 서열정보 검색 시스템 개발
가. ProSeS (Protein Sequence Search 시스템 : 색인기반 단백질서열 분석 시스템) 구축
- 종합적인 서열정보 분석 지향
- 단백질아미노산 서열의 n-gram 색인기법을 통한 정보검색기반 서열 검색 시스템 구축
- 단백질서열 검색결과에 대한 주제어 제시 등과 같은 데이터 마이닝 기법 개발
- 신규 단백질의 기능/구조적 분류, 세포내 위치 예측 결과 등 제공

나. ProNGF (Protein N-Gram Frequency : 단백질 N-Gram 빈도 데이터베이스) 구축
- 단백질서열 내 N-Gram빈도를 색인화한 단백질 서열 N-Gram빈도 데이터베이스 및 검색서비스 구축

3) 생명정보검색시스템 기반의 단백질 서열 분류 시스템 개발
가. ProSLP (Protein Subcellular Localization Prediction : 단백질 세포내 위치 예측) 데이터베이스 구축
- 색인기반 단백질 세포내 위치 예측 시스템으로서의 ProSLP 웹 서비스 시스템 개발
- CELL_LOC 기반의 세포내 위치 데이터 집합 구축
- SWISS-PROT 기반의 세포내 위치 데이터 집합 구축

나. ProFaC (Protein Family Classification : 단백질서열 기능분류 서비스 시스템 개발
- iProClass의 기능성 단백질분류인 superfamily 분류체계에 따른 기능 분류서비스 시스템 구축
- PFam 분류체계에 따른 단백질의 2차원 구조 분류시스템 개발
- InterPro 및 BLOCKS의 분류체계 검토 및 분류서비스 타당성 검토

4) 신호전달 데이터베이스 구축 및 네비게이션 시스템 개발
가. 신호전달 통합 데이터베이스 구축 (1차년도)
- 국외 공공 신호전달 DB 수집 및 통합 DB 구축
- 기존 구축된 Genome DB, Protein DB 등과 신호전달 DB와 연계

나. Signal Transduction Network DB 구축 (2차년도)
- 통합 DB를 이용한 Signal Transduction Network DB 구축
- 단백질간의 연관성을 표현하는 STML(Signal Transduction Markup Language)개발

다. 신호전달 네비게이션 시스템 개발 (3차년도)
- 신호전달 DB 가시화 시스템 개발

2. 생명정보 활용체제 구축

1) 생명정보 DB 구축 및 인터페이스 개발
- 생명정보 DB 구축 및 유지보수
- 웹 서비시와 연계를 위한 DB 질의 튜닝 및 스키마 튜닝
- 자동 업데이트 모듈 개발적용, DB 저장, 적제, 검색 등 인터페이스 개발

2) 생명정보 웹 서비스 체제 구축 연구

- UDDI 레지스트리 구축
- 생명정보 서비스 등록, 서비스 검색 개발
- 생명정보 분류 체계 작성
- 포탈 서비스 시범 사이트 구축(통합 검색, 분석)

3. 생명정보 지원체제 구축

1) 생명정보 시스템 운영체제 구축
- 생명정보시스템 도입/운영/관리
- 서버 보안시스템 및 방화벽 구축
- 클러스터 원격배치서비스 시스템 구축

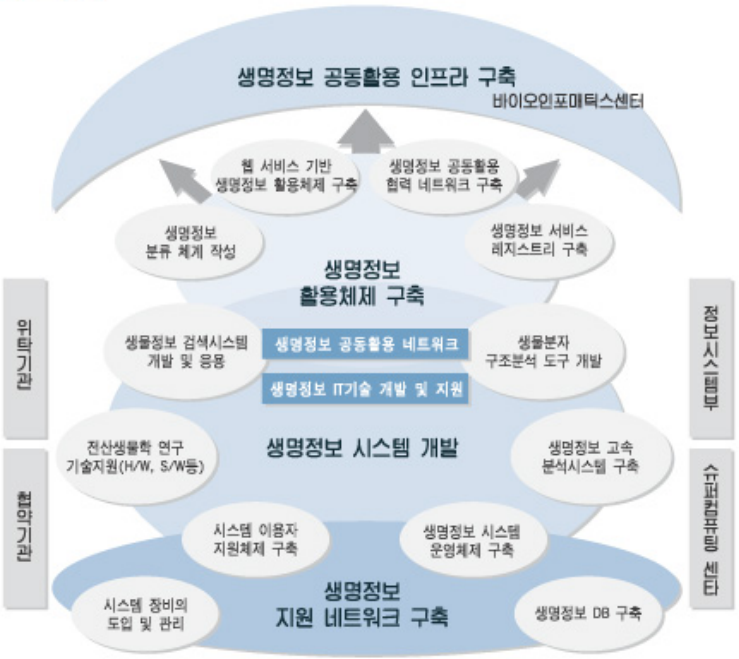
2) 생명정보시스템 이용자 지원체제 구축
- 이용자 지원 정책 수립
- 사용자 요구조사 및 이용 만족도 조사
- 유관기관 협력 네트워크 구축

3) 생명정보 컨텐츠 구축
- 전산생물학(화학) 컴퓨팅 자원 및 기술 지원
- 바이오인포매틱스 최신동향 분석 및 바이오인포매틱스 용어사전 구축
- FTP 사이트 증설 및 최신성 유지

4) 전산생물학 연구기술 지원(하드웨어/소프트웨어)

5) 생명정보 공동 활용 협력 네트워크 구축

Ⅳ. 추진체계



V. 기대효과

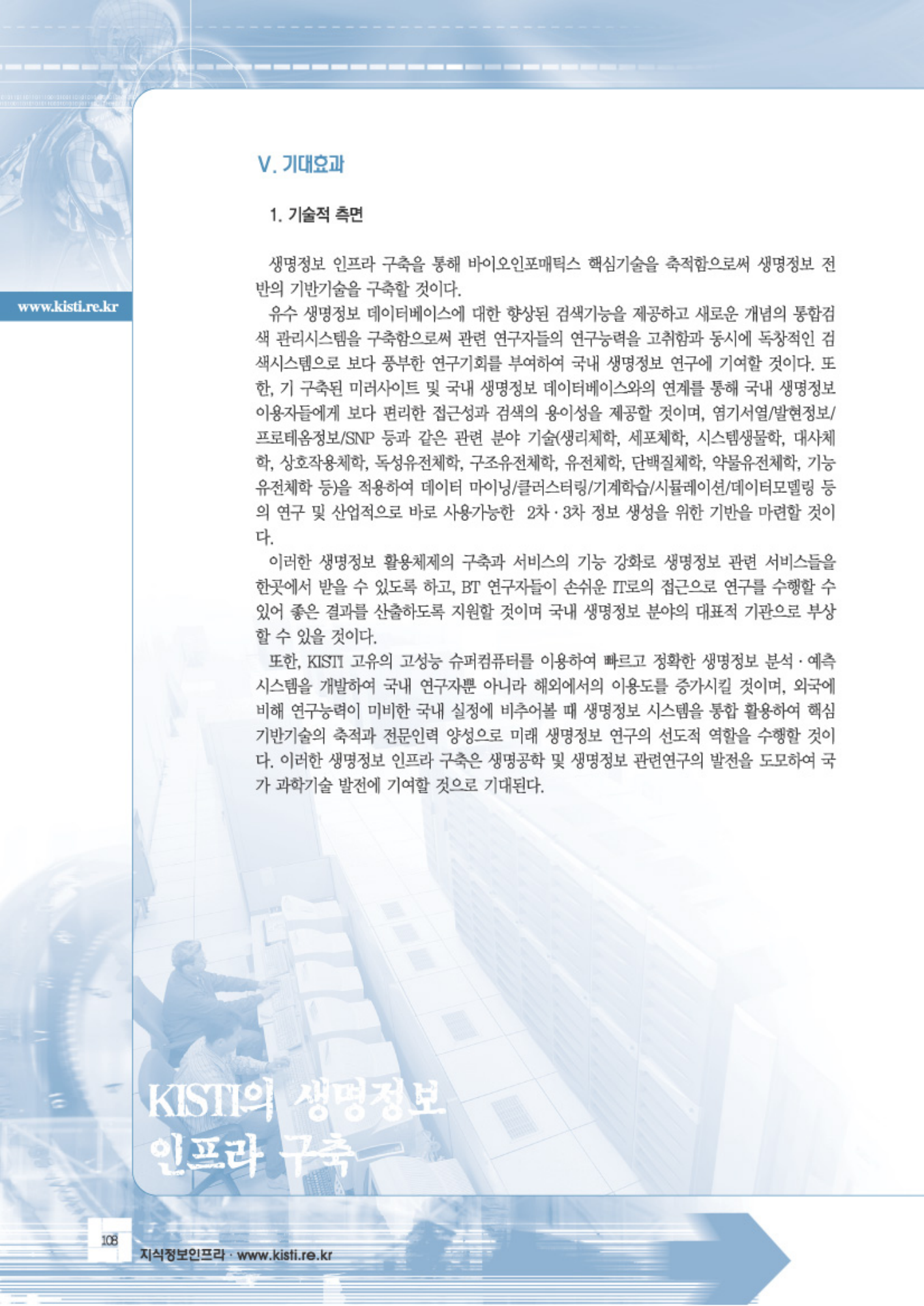
1. 기술적 측면

생명정보 인프라 구축을 통해 바이오인포매틱스 핵심기술을 축적함으로써 생명정보 전반의 기반기술을 구축할 것이다.

유수 생명정보 데이터베이스에 대한 향상된 검색기능을 제공하고 새로운 개념의 통합검색 관리시스템을 구축함으로써 관련 연구자들의 연구능력을 고취함과 동시에 독창적인 검색시스템으로 보다 풍부한 연구기회를 부여하여 국내 생명정보 연구에 기여할 것이다. 또한, 기 구축된 미러사이트 및 국내 생명정보 데이터베이스와의 연계를 통해 국내 생명정보 이용자들에게 보다 편리한 접근성과 검색의 용이성을 제공할 것이며, 염기서열/발현정보/프로테옴정보/SNP 등과 같은 관련 분야 기술(생리체학, 세포체학, 시스템생물학, 대사체학, 상호작용체학, 독성유전체학, 구조유전체학, 유전체학, 단백질체학, 약물유전체학, 기능유전체학 등)을 적용하여 데이터 마이닝/클러스터링/기계학습/시뮬레이션/데이터모델링 등의 연구 및 산업적으로 바로 사용가능한 2차 · 3차 정보 생성을 위한 기반을 마련할 것이다.

이러한 생명정보 활용체제의 구축과 서비스의 기능 강화로 생명정보 관련 서비스들을 한곳에서 받을 수 있도록 하고, BT 연구자들이 손쉬운 IT로의 접근으로 연구를 수행할 수 있어 좋은 결과를 산출하도록 지원할 것이며 국내 생명정보 분야의 대표적 기관으로 부상할 수 있을 것이다.

또한, KISTI 고유의 고성능 슈퍼컴퓨터를 이용하여 빠르고 정확한 생명정보 분석 · 예측 시스템을 개발하여 국내 연구자뿐 아니라 해외에서의 이용도를 증가시킬 것이며, 외국에 비해 연구능력이 미비한 국내 실정에 비추어볼 때 생명정보 시스템을 통합 활용하여 핵심 기반기술의 축적과 전문인력 양성으로 미래 생명정보 연구의 선도적 역할을 수행할 것이다. 이러한 생명정보 인프라 구축은 생명공학 및 생명정보 관련연구의 발전을 도모하여 국가 과학기술 발전에 기여할 것으로 기대된다.



2. 경제 · 산업적 측면

생명정보 인프라 구축 및 생명정보 통합시스템의 개발은 의약, 화학, 환경, 농업 등의 관련 연구 및 산업의 활성화에 기여할 수 있다. 분산되어 있는 유전정보 자원관리와 활용시스템 구축을 통해 IT와 BT의 융합을 꾀하고 기술력 축적과 이를 통한 국내 생명공학 및 관련 산업의 국제 경쟁력을 제고시킬 것이다.

최근 유용단백질의 생산이 빠른 속도로 증가하고 있으며, 생명현상에 중요한 역할을 하는 유전자 관련 특허가 출원되는 추세이므로, 이에 생명정보 시스템이 이용된다면 국외의 다른 연구자들보다 새로운 유전자 또는 Drug Target 발견에 앞설 수 있어 의약과 같은 관련 산업에 막대한 경제적 기여가 가능할 것임이 분명하다. 이러한 생명정보 인프라 구축은 생명정보학 연구자들이나 신약개발 관련 연구자들 및 국외의 생명정보 분야 분석 검색시스템을 이용하는 국내 산 · 학 · 연 연구자들에게 관련정보 및 기반기술을 제공하여 연구의 효율성을 높이고, 생명과학 기술 산업의 국제 경쟁력 확보와 고부가가치의 경제효과 창출 및 신약 개발을 위한 핵심 기술 개발을 통해 의학 · 약학 등의 의료분야 발전에도 크게 기여할 것이다.

최근 전 세계적으로 폭발적으로 증가하고 있는 공공 생명정보 데이터를 해외 서비스 기관에 의존하지 않고 국내 독자적 서비스를 가능케 할 수 있어 경제적 효과를 거둘 것이며, 초고속 네트워크와 슈퍼컴퓨터를 연계하여 대용량의 초고속 고부가 정보검색 및 분석이 가능케 될 것이다. 또한, 기하급수적으로 증가하고 있는 생명정보를 효율적으로 처리할 수 있는 검색시스템의 개발로 BT 연구자들의 시간과 노력을 절감시키고 산 · 학 · 연 각 연구실에서 필요한 데이터베이스 및 소프트웨어, 하드웨어 자원을 제공함으로써 국내에서의 중복 투자를 방지할 수 있으며 구축된 생명정보 시스템을 국가유전체정보센터 인프라와 유기적으로 결합하여 시너지 효과를 창출할 수 있을 것이다.